The University of Warwick

THEORY OF COMPUTATION

REPORT NO.37

GAUSS' ALGORITHM FOR THE SOLUTION OF QUADRATIC DIOPHANTINE EQUATIONS

BY

MEURIG BEYNON AND COSTAS ILIOPOULOS

Department of Computer Science
University of Warwick
COVENTRY CV4 7AL
ENGLAND.

May 1981

Gauss' algorithm for the solution of quadratic Diophantine equations

by

Meurig Beynon and Costas Iliopoulos
Department of Computer Science
University of Warwick
COVENTRY CV4 7AL
ENGLAND

To Megan & Granville and Despina & Spiros

Contents.

§1. Introduction.

An algorithmic solution to the problem of deciding whether a given integer is representable by a given binary quadratic form was first described by Gauss in Disquisitiones Arithmeticae (1801) (reference [G] below). This solution forms a part of the elementary theory of quadratic forms as developed by Gauss in Articles 153-222 of [G].

The study of such classical algorithmic problems in number theory has recently acquired topical and practical interest as a result of the development of the RSA public-key cryptosystem, which depends upon the apparent intractability of factorising large integers. The aim of this report is to present the essential concepts and results used in the description and justification of Gauss's algorithm in a form which assumes only elementary number-theory, and is accessible to the reader whose specialist interest is in algorithms. For the most part, the exposition is closely based upon Gauss' original account, with a few minor simplifications. The most novel (and possibly idiosyncratic!) feature of the presentation is the treatment of equivalence of forms having positive non-quadratic determinant (§8), which is based upon results of [B].

§2. Preliminaries on quadratic forms.

A binary quadratic form is a polynomial of the form

$$F(x,y) = ax^2 + 2bxy + cy^2,$$

where a,b and c are integers. The notation F=(a,b,c) will be used to denote such a form, and "form" will be used as a synonym for "binary quadratic form".

The matrix of the form F=(a,b,c) is the 2x2 matrix

$$\begin{pmatrix} a & b \\ b & c \end{pmatrix},$$

and the determinant of the form F is $D = b^2-ac$. (Note that this differs in sign from the determinant of the matrix of F.)

If F=(a,b,c), and p and q are integers such that F(p,q)=K, then (p,q) is "a representation of the integer K by the form F".

Equivalently, x=(p,q) is a representation of K by the form F with matrix M if $xMx^T = K$.

Let F and G be forms with associated matrices M and N respectively. F and G are equivalent forms if there is a 2x2 integer matrix X of determinant 1 such that $X^TMX = N$.

Note that if the forms F and G are equivalent, they necessarily represent the same set of integers, and have the same determinant.

Notation:

The notation [x,y,z,t] will be used to denote the 2x2 matrix:

$$\begin{pmatrix} x & y \\ z & t \end{pmatrix}.$$

As in Gauss [G], a "perfect square" is called a "quadratic integer".

If F=(a,b,c), then ||F|| will denote $\max(|a|,|b|,|c|)$.

§3. An overview of the algorithm:

Gauss [G] describes algorithms to solve the following related problems:

Problem 1:
Given integers a,b,c and K, determine whether K is representable by the form F=(a,b,c), and, if so, find all such representations of K.

Equivalently, determine whether the equation

$$ax^2 + 2bxy + cy^2 = K \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (E1)$$

has an integral solution (x,y), and, if so, find all such pairs (x,y).

Problem 2:
Given integers a,b,c,d,e and f, determine whether the equation

$$ax^2 + 2bxy + cy^2 + 2dx + 2ey + f = 0 \quad \ldots\ldots\ldots \quad (E2)$$

has an integral solution (x,y), and, if so, find all such pairs (x,y).

The determinant $D = b^2 - ac$ determines the algorithms to be used. The case D=0 is exceptional, and special techniques apply. (See §4 and Art.'s 215, 219, and 220). Provided that D is non-zero, Problem 2 can be reduced to Problem 1 (see Art.216 and §10). For this reduction, an effective procedure for identifying all solutions (if any) of (E2) is required when the associated instance of (E1) has infinitely many solutions i.e. if D is a non-quadratic positive integer (Art.217,221), or D is quadratic and K=0 (Art.201).

A form F can represent 0 only if it has quadratic or zero determinant, and this case requires a special method (see §9 and Art.212). If D and K are both non-zero, Gauss' algorithms for the solution of Problem 1 share common principles, and are based on the following strategy:

(1) determine whether D is a quadratic residue modulo K. If so, find all square roots of D modulo K.

(2) For each square-root V of D modulo K, determine whether the forms F=(a,b,c) and $G=(K,V,(V^2-D)/K)$ are equivalent. If so, determine all equivalences between F and G.

The justification for this strategy appears in §5; it depends upon two fundamental results:

(1) (D/K)=1 is a necessary condition for K to be representable by the form F, and each representation is associated with a particular square-root V of D modulo K (Art.154-5).

(2) Equivalence between F and G is a necessary and sufficient condition for the form F to represent K, and the set of representations is in canonical 1-1 correspondence with the set of such equivalences (Art.168-9).

If one equivalence between a pair of forms F and G is known, the problem of determining all equivalences between F and G can be resolved by describing all equivalences between F and itself. If F=(a,b,c), and m = gcd(a,2b,c), then such equivalences are in canonical 1-1 correspondence with integer pairs (t,u) such that:

$$t^2 - Du^2 = m^2.$$

This correspondence is described in §6, which is based on Art.162.

For non-zero D, the problem of deciding when two forms of determinant D are equivalent, and determining such an equivalence explicitly when appropriate, is solved by techniques which depend upon the nature of D. There are three cases, corresponding to D<0 (§7 and Art.171-9), D>0 and non-quadratic (§8 and Art.183-200), and D>0 quadratic (§9 and Art.206-10). Similar principles are used in each of these three cases: the identification of a special class of `reduced' forms, a `reduction algorithm' for deriving a reduced form equivalent to a given form, an effective criterion for equivalence of reduced forms, and an effective procedure for explicitly describing equivalences between reduced forms.

§4. The case of zero determinant.

Theorem 4.1:
Let F=(a,b,c) be a form with zero determinant, and suppose that m=gcd(a,2b,c). Then there are co-prime integers g and h such that:

$$ax^2 + 2bxy + cy^2 = m(gx + hy)^2 .$$

Proof:
Since $b^2 = ac$, any common divisor of a and c divides b, so that $m = \gcd(a,c)$, and $(b/m)^2 = (a/m)(c/m)$ where a/m and c/m are co-prime. But a/m and c/m are then necessarily quadratic integers, and $a/m=g^2$, $c/m=h^2$, where g and h are co-prime. Moreover, with appropriate choice of sign for g:

$$ax^2 + 2bxy + cy^2 = m(g^2x^2 + 2ghxy + h^2y^2) = m(gx+hy)^2 .$$

The next corollary follows immediately from Thm.4.1:

Cor. 4.1.1:
Let F=(a,b,c) be a form of determinant zero, and let m,g and h be as in Thm.4.1.

(1) The integer K is representable by F if and only if the quotient K/m is a quadratic integer n^2.

and (2) if K is representable by F, and n is as in (1), the representations of K by F are the integer pairs (p,q) such that $pg+qh=\pm n$.

To solve Problem 1 completely when F has zero determinant, it remains to describe the set of solutions referenced in Cor.4.1.1(2) explicitly:

Lemma 4.2:
Suppose that g and h are co-prime integers, and that G and H are integers such that Gg+Hh = 1. If d is a given integer, the general solution of the equation gx+hy = d is the set of integer pairs:

$$(x(k),y(k)) = (Gd+hk,Hd-gk)$$

where k assumes any integer value.

Proof:
It is easy to verify that (x(k),y(k)) is a solution for each integer k. But if (X,Y) is a solution, then GgX+GhY = Gd, whence X = Gd+h(HX-GY). Similarly, Y = Hd-g(HX-GY), so that each solution has the form (x(k),y(k)).

The solution to Problem 2 subject to $b^2 - ac = \emptyset$ will now be described.
By Thm.4.1, co-prime integers g and h may be chosen so that, upon substitution of z=gx+hy, equation (E2) becomes:

$$mz^2 + 2dx + 2ey + f = \emptyset \ldots\ldots\ldots \text{(E3)}.$$

There are then two cases:

Case 1: $dh \neq eg$
Since h is non-zero, equation (E3) is equivalent to:
$hmz^2 + 2dhx + 2ehy + fh = \emptyset$, or $hmz^2 + 2dhx + 2ez - 2egx + fh = \emptyset$.
which in turn is equivalent to $x = (hmz^2+2ez+fh)/2(eg-dh)$. A similar derivation now shows that (x,y) is a solution of (E2) iff

$$y = (gmz^2 + 2dz + fg) / 2(dh - eg),$$

whence the general solution of (E2) in this case is the set of pairs of the form (x(k),y(k)), where

$$x(k) = (hmk^2+2ek+fh)/2(eg-dh) \text{ and } y(k) = (gmk^2+2dk+fg)/2(dh-eg).$$

and k is an integer for which both x(k) and y(k) are also integral.
Since the integral nature of x(k) and y(k) depends only upon the residue class of k modulo $2|eg-dh|$, this form of solution is sufficient for effectively enumerating all solutions of (E2), and effectively deciding whether at least one solution exists.

Case 2: $dh = eg$

Since g and h are co-prime, $d/g = e/h$ is an integer k. Equation (E3) is then equivalent to

$$mz^2 + 2dx + 2ey + f = 0 \text{ or } mz^2 + 2k(gx+hy) + f = 0.$$

Thus, an integer solution of (E2) is possible only if

$$mz^2 + kz + f = 0$$

has an integral root; that is, only if there is a positive integer r such that $r^2 = k^2 - mf$ and m divides $-k+r$ or $-k-r$. If p is such an integral root, the appropriate values of x and y satisfying (E2) may then be obtained from Lemma 4.2 by solving $gx+hy = p$.

Notes on §4.

Theorem 4.1 and Cor.4.1.1 follow [G] Art.215. The solution to Problem 2 is described in Art.'s 219 and 220.

The explicit solution of Problem 1 in this case can be determined in time polynomial in $\log(\|F\|)$. As described, the explicit solution of Problem 2 requires consideration of all residues modulo $|eg-dh|$, and cannot be determined in time polynomial in $\log(\max(|a|,|b|,|c|,|d|,|e|,|f|))$.

§5. Generalities on representations of integers by non-singular forms.

The results presented in this section are fundamental to the analysis of representations of non-zero integers by non-singular forms. (The representation of zero is the only instance of Problem 1 which requires exceptional treatment, but such representation is possible only if

$$ax^2 + 2bxy + cy^2 = 0$$

whence $(ax+by)^2 = Dy^2$, and D is quadratic (see §9)). In the context of Problem 1, it is convenient to consider only "primitive" representations, that is, representations (m,n) such that m and n are co-prime. There is no essential loss of generality, provided that the identification of quadratic factors of K provides no obstacle. (If (P,Q) is a representation of the integer K by the form F, and d=gcd(P,Q), then d^2 divides K, and (p=P/d,q=Q/d) is a representation of K/d^2 by F, such that p and q are co-prime. Moreover, all representations of K by F arise in this manner.) In the sequel, it is to be understood (unless explicitly stated otherwise) that `a representation (m,n) of an integer K by a form of determinant D' entails gcd(m,n)=1 and both K and D non-zero.

Theorem 5.1:

Let F=(a,b,c) be a form of non-zero determinant D, K a non-zero integer, and (p,q) a representation of K by F. Let C be the set of integer pairs (P,Q) such that Pp + Qq = 1.

Then

(1) V(P,Q) = P(bp+cq) - Q(ap+bq) is independent modulo K of the choice of (P,Q) in C.

(2) if (P,Q) is in C, then V(P,Q) is a square-root of D modulo K.

(3) the range of V, considered as an integer-valued function on C, is a residue class modulo K.

Proof:

(1) By Lemma 4.2, C is the set of pairs:

$$(P(k),Q(k)) = (P+qk,Q-pk)$$

where k is an integer. Thus, values of V which correspond to different elements of C differ by qk(bp+cq)+pk(ap+bq) = kK, where k is integral.

(2) $(v(P,Q))^2 - D.(Pp+Qq)^2 = (ap^2 + 2bpq + cq^2)(aP^2 - 2bPQ + cQ^2)$
$= K.(aP^2 - 2bPQ + cQ^2)$.

(3) V(P(k),Q(k)) - V(P,Q) = kK, and k is an arbitrary integer.

Cor. 5.1.1:

If the form Q of determinant D represents the integer K, then (D/K)=1.

Definition:

The representation (p,q) of K by the form F of determinant D is said to "belong to the value V" where V=V(P,Q) is as defined in Thm.5.1.

Theorem 5.2:

Let F be a form of determinant D, and K a non-zero integer such that (D/K)=1. For each square-root V of D modulo K there is a canonical 1-1 correspondence between:

A: the set of representations (p,q) of K by F belonging to the value V, and B: the set of equivalences between F and the form $G=(K,V,(V^2-D)/K)$.

Under this correspondence, the representation (p,q) of K by F belonging to the value V is associated with the equivalence from F to G defined by the matrix [p,-Q,q,P], where P and Q satisfy

$$Pp + Qq = 1 \text{ and } P(bp+cq)-Q(ap+bq) = V.$$

Proof:

(p,q) is a representation of K by F which belongs to the value V
iff $ap^2 + 2bpq + cq^2 = K$, and there are integers P and Q such that:
$Pp + Qq = 1$ and $P(bp+cq)-Q(ap+bq) = V$ (by Thm.5.1 (3))
iff Y = [p,-Q,q,P] defines an equivalence between F and G.

The correspondence defined in this way is 1-1, since P and Q are uniquely determined subject to Pp+Qq=1 and V(P,Q)=V.

Notes on §5.

Theorems 5.1 and 5.2 are based on [G] Art.'s 154-6 and 167-9 respectively.

The problem of identifying square factors of K is apparently no easier than the general factorisation problem. If the complete factorisation of K is known, and a quadratic non-residue for each distinct prime divisor of K is supplied, then a root V of D modulo K can be determined in $O(M(\log(|K|))(\log(|K|))^2)$ elementary operations using the Tonelli-Shanks algorithm [L1]. The number of square-roots of D modulo K can be large, however; there are two square-roots of D modulo each distinct prime power dividing K, and hence 2^r square-roots of D modulo K, where r (of order $\log(K)/\log(\log(K))$ is the number of distinct prime factors of K.

These complexity considerations are relevant to all the algorithms based on the principles of Theorems 5.1 and 5.2.

§6. On determining all proper equivalences between two forms.

Theorem 5.2 indicates the importance of being able to determine all proper equivalences between a given pair of forms having the same determinant. If (F,G) is such a pair of forms, it is convenient to consider the problem of finding all equivalences from F to G as comprising two sub-problems:

(1) finding a single equivalence between F and G

and (2) determining all equivalences between F and itself.

The solution of (1) will be developed in later sections in conjunction with algorithms for deciding equivalence of forms with the same determinant D. The techniques used differ according to the nature of D. In this section, a general approach to the solution of (2) is described.

Theorem 6.1:

Let $F=(a,b,c)$ be a form, M the matrix of F, and $X = [x,y,z,t]$ an integer matrix such that $\det X = 1$.

Then $X^TMX = M$ if and only if

(A) $cz = -ay$

(B) $a(t-x) = 2bz$

and (C) $c(x-t) = 2by$.

Moreover, (A),(B) and (C) together imply

(P) $(ax+bz)^2 - Dz^2 = a^2$

(Q) $(by+ct)^2 - Dy^2 = c^2$

and (R) $(x+t)^2 b^2 - D(x-t)^2 = 4b^2$.

Proof:

$X^TMX = M \iff X^TM = MX^{-1} \iff$ (A),(B) and (C) hold, by direct computation.

When (A),(B) and (C) hold:

$$(ax+bz)^2 - Dz^2 = a^2x^2 + 2abxz + acz^2$$
$$= a^2x^2 + a^2x(t-x) - a^2yz \text{ by (A) and (B)}$$
$$= a^2 \text{ since } xt-yz = 1.$$

This establishes (P), and a symmetrical argument proves (Q). Also:

$$(x+t)^2 b^2 - D(x-t)^2 = 4b^2xt + ac(x-t)^2 = 4b^2(xt-yz) \text{ by (B) and (C)}.$$

Cor.6.1.1:

Suppose that M and X are as in Thm.6.1, and that $X^TMX = M$. Let p,q and r be any integers, and define

$m = pa + 2qb + rc$

$T = p(ax+bz) + qb(x+t) + r(ct+by)$

and $U = -pz + q(x-t) + ry$.

Then (a) $mz = -aU$

(b) $my = cU$.

(c) $m(x-t) = 2bU$

(d) $m(x+t) = 2T$

(e) $m(ax+bz) = aT$

and (f) $m(ct+by) = cT$.

Moreover, (a),(b) and (c) together imply (A),(B) and (C) of Thm.6.1.

Proof:

(a) $mz = paz + 2qbz + rcz = -aU$ by Thm.6.1 (A) and (B).

(b) $m(x+t) = (pa + 2qb + rc)(x+t)$
$= p(ax + at) + 2qb(x + t) + r(cx + ct)$
$= 2T$ by Thm.6.1 (B) and (C).

The proofs of (c) and (d) are similar.

(e) $m(ax + bz) = p(a^2x+abz) + q(2abx+2b^2z) + r(acx+bcz)$
$= ap(ax+bz) + aq(2bx+bt-bx) + ar(cx-by)$
by Thm.6.1 (B) and (A)
$= aT$ since $cx-by = ct+by$ by Thm.6.1 (C).

The proof of (f) is similar.

(a) and (b) $\Rightarrow mcz = -acU = -amy \Rightarrow$ (A), since m is non-zero.

Similarly: (a) and (c) $\Rightarrow ma(t-x) = -2abU = 2bmz \Rightarrow$ (B),

whilst (b) and (c) $\Rightarrow mc(x-t) = 2bcU = 2bmy \Rightarrow$ (C).

Cor.6.1.2: Under the hypotheses of Cor.6.1.1:
(1) T and U depend upon m, but not the particular choice of p,q and r and (2) $T^2 - DU^2 = m^2$.

Proof:
(1) is implicit in the identities (a),(b),(c) and (d), since at least one of a,b and c is non-zero.
(2) There are three cases to consider, according as a,b or c is non-zero.
If $b \neq 0$, then Thm.6.1 (R) and Cor.6.1.1 (c) and (d) show that
$4b^2(T^2 - DU^2) = 4b^2 m^2$, whence $T^2 - DU^2 = m^2$.
If $a \neq 0$, then Thm.6.1 (P) and Cor.6.1.1 (a) and (e) show that
$a^2(T^2 - DU^2) = a^2 m^2$, whence $T^2 - DU^2 = m^2$.
The case $c \neq 0$ is dealt with similarly.

Cor.6.1.3:
Suppose that F and M are as in Thm.6.1, and that m = gcd(a,2b,c). There is a canonical 1-1 correspondence between
R: the set of integer matrices of determinant 1 such that $X^TMX = M$, and S: the set of solutions (T,U) of the equation $T^2 - DU^2 = m^2$.

Proof:
Let X be in R, and let p,q and r be integers such that m = pa+qb+rc. By Cor.'s 6.1.1 and 6.1.2, there are integers T and U, independent of the particular choice of p,q and r, such that (T,U) is in S.
Now suppose that (T,U) is a solution of $T^2 - DU^2 = m^2$, and define:
$x = (T+bU)/m$, $y = cU/m$, $z = -aU/m$ and $t = (T-bU)/m$.
It easy to verify that x,y,z and t are rational numbers which satisfy conditions (a) through (c) of Cor.6.1.1, and they accordingly satisfy conditions (A) through (C) of Thm.6.1. Moreover:
$$\det X = (T^2 - (b^2 - ac)U^2)/m^2 = 1,$$
so that X is seen to be in R provided only that x,y,z and t are integral.
Both y and z are integers, since m divides a and c. Since m divides 2b, it divides both T+bU and T-bU if it divides one or other. But
$T^2 - DU^2 = m^2$, whence m^2 divides $T^2 - b^2U^2 = (T-bU)(T+bU)$.
The mappings between R and S defined in this way are inverse bijections, and define a 1-1 correspondence between R and S.

Cor.6.1.4:
Let F=(a,b,c) and G be forms, and let M and N be their respective matrices. If Y = [p,q,r,s] is an integer matrix of determinant 1 such that $Y^TMY = N$, then every matrix
$$[\ (pT+(pb+rc)U)/m,\ (qT+(qb+sc)U)/m,\ (rT-(rb+pa)U)/m,\ (sT-(qa+sb)U)/m\ ]$$
where m = gcd(a,2b,c), and (T,U) satisfies $T^2 - DU^2 = m^2$, also defines an equivalence from F to G, and all such equivalences arise in this way.

Proof:
Suppose that Z defines an equivalence from F to G, so that $Z^TMZ = N$. Then $Z^TMZ = Y^TMY$, whence $(ZY^{-1})^TM(ZY^{-1}) = M$, and by Cor.6.1.3:
$$ZY^{-1} = X = [\ (T+bU)/m,\ cU/m,\ -aU/m,\ (T-bU)/m\ ]$$
where $T^2 - DU^2 = m^2$. Thus Z = XY, which has the stated form.
Conversely, any matrix having the form XY defines an equivalence between F and G, since $(XY)^TMXY = Y^TX^TMXY = Y^TMY = N$.

The next corollary is a direct consequence of Cor.6.1.4 and Thm.5.2:

Cor.6.1.5:
Let (p,q) be a representation of K by the form F=(a,b,c) which belongs to the value V. If m=gcd(a,2b,c), and t and u are integer solutions of the equation $t^2 - Du^2 = m^2$, then:
$$P = (pt + (pb+qc)u)/m \quad \text{and} \quad Q = (qt - (qb+pa)u)/m$$
also defines a representation (P,Q) of K by F belonging to V, and all such representations arise in this way.

Notes on §6.

The essential results of this section are to be found in [G] Art.162, but the proofs have been simplified in minor respects.

§7. The case of negative determinant.

Throughout this section, D will denote a positive integer.

Definition:
A form $F = (a,b,c)$ of determinant $-D$ is reduced if
$$2|b| \leqslant |a| \leqslant |c|.$$

If F is reduced, then $D+b^2 = ac \geqslant a^2$, whilst $0 \leqslant |b| \leqslant |a|/2$ entails $b^2 \leqslant a^2/4$, so that $|a| \leqslant \sqrt{(4/3).D}$.

Theorem 7.1:
If $F = (a,b,c)$ is a form of determinant $-D$, then F is properly equivalent to a reduced form.

Proof:
If F is not itself reduced, then consider the sequence of forms
$$F = F_0, F_1, \ldots, F_t,$$
where $F_i = (a_i,b_i,c_i)$, and F is defined in such a way that
$$a_{i+1} = c_i \text{ and } b_{i+1} = -b_i + k_i c_i,$$
where $|b_{i+1}| \leqslant |c_i|/2$. (Note that this can exceptionally permit two choices of k_i, if c_i is even, and $-b_i \equiv c_i/2 \pmod{c_i}$). The sequence a consists of positive integers (since $-D = b^2-ac$), and there must be a first integer $k_i \geqslant 1$ for which $|a_K| \leqslant |a_{K+1}|$. The form $F_K = (a_K,b_K,c_K)$ is then reduced. Explicitly: $|a_K| \leqslant |c_K| = |a_{K+1}|$, and $|b_K| \leqslant |c_{K-1}|/2$ by definition, whence
$$2|b_K| \leqslant |a_K| = |c_{K-1}| \leqslant |c_K|.$$

Theorem 7.2:
Let $F=(a,b,c)$ and $G=(A,B,C)$ be two reduced forms of determinant $-D$. Then F and G are equivalent if and only if
at least one of the following conditions holds:
(c1) F=G
(c2) a=A, c=C, b=-B and $a=\pm 2b$,
(c3) a=A, c=C, b=-B, and a=A=c=C.

Proof:
Suppose that the matrix $X=[x,y,z,t]$ defines a proper equivalence between F and G. The identity $X^TF = GX^{-1}$ is equivalent to the equations:
$$\begin{aligned} At - ax &= (B+b)z \ldots (e1),\\ cz + Ay &= (B-b)x \ldots (e2),\\ ay + Cz &= (B-b)t \ldots (e3),\\ \text{and } Cx - ct &= (B+b)y \ldots (e4),\end{aligned}$$
where $xt-yz = 1$, and (without loss of generality) $|A| \geqslant |a|$.
Because F and G have negative determinant, a has the same sign as c, and A the same sign as C. From (e1) and (e3) above:
$$a = a(xt-yz) = At^2 + Cz^2 - 2Bzt \ldots (e5),$$
whence $aA = (At-Bz)^2 + Dz^2 > 0$, and a,A,c and C all have the same sign. Re-arranging (e5) shows that $At^2 + Cz^2 - a = 2Bzt$, so that
$$\begin{aligned} |2Bzt| = |At^2 + Cz^2 - a| &\geqslant \min(|A|,|C|)(z^2+t^2) - |a| \ldots\ldots\ldots (i)\\ &\geqslant |A|(z^2+t^2) - |A| \ldots\ldots\ldots\ldots\ldots (ii)\\ &= |A|(z^2+t^2-1)\\ &\geqslant 2|B|(z^2+t^2-1)\end{aligned}$$
using the fact that the form G is reduced. But now
$$z^2 + t^2 - |zt| = (|z| - |t|)^2 + |zt| \geqslant 1,$$
since z and t are not both zero, so that the above inequalities are only consistent if $(|z| - |t|)^2 + |zt| = 1$, and none of the other 'inequalities' is strict. Thus the only possible values for $|z|$ and $|t|$ are 0 and 1, and A and a are necessarily equal, since inequality (ii) cannot be strict.
There are then two cases:

Case 1: $|z| = 1$.

In this case, C=A since inequality (i) cannot be strict. By (e1):

$$A(t-x) = a(t-x) = (B+b)z.$$

If $x=t$ then $B = -b$, whilst otherwise

$$|B+b| = |A||t-x| \geq |A| = |a| \geq 2.\max(|b|,|B|),$$

which entails $|B| = |b|$. But A=a and $b=\pm B$ is sufficient to ensure that c and C have the same modulus, and are thus equal, whence A=a=C=c and $b=\pm B$, and (c1) or (c3) applies.

Case 2: $|z| = \emptyset$.

In this case, $xt = 1$, so that $x = t = \pm 1$. If $y=\emptyset$, then F=G; otherwise equation (e2) above reduces to the form $ay = (B-b)x$, whence

$$|a| \leq |a||y| = |B-b| \leq 2.\max(|B|,|b|) \leq |A| = |a|.$$

This is consistent only if $B = -b$, and $2.|B| = 2.|b| = |a| = |A|$. As in Case 1, this entails C=c, and (c2) applies.

Theorem 7.2 solves the problem of identifying equivalent forms of determinant -D, and finding an explicit equivalence between a pair of equivalent forms.

Suppose that F=(a,b,c) is a form of determinant -D with matrix M, and that G is an equivalent form. Let $m = \gcd(a,b,c)$. By Cor.6.1.3, the problem of determining all equivalences between F and G can be solved by identifying all matrices X such that

$$X^TMX = M \quad \ldots\ldots\ldots\ldots \text{(p1)}$$

or equivalently, finding all integer solutions (T,U) of

$$T^2 + DU^2 = m^2 \quad \ldots\ldots\ldots \text{(p2)}.$$

As Cor.6.1.1 (d) shows, a matrix X satisfying (p1) corresponds canonically to a solution (T,U) of (p2) where 2T/m is the trace of X. If X satisfies (p1), then so also does X^k, for any integer k. The set of solutions of (p2) is patently finite, so that the set of traces of matrices of the form X^k is also finite. But if X=[x,y,z,t], then the trace of X^2 is

$$tr(X^2) = x^2+t^2+2yz = (tr(X))^2 - 2,$$

since X has determinant 1. Hence tr(X) has modulus at most 2 (consider the sequence of matrices $X, X^2, X^4, X^8, \ldots$).

The above argument shows that a necessary condition for (T,U) to be a solution of (p2) is that $T = \pm m, \pm m/2$ or $\emptyset$. Whatever the value of D, there is a solution $(\pm m, \emptyset)$. If (p2) has a solution of the form $(\pm m/2, U)$, then $4DU^2/m^2=3$; since $4D/m^2$ is an integer, this is possible only if $U=\pm 1$, and D is of the form $3m^2/4$. If $(\emptyset,U)$ is a solution of (p2), then $DU^2=m^2$, and $(4D/m^2).(U/2)^2=1$. A priori, this is consistent with

$$D=m^2 \text{ and } U=\pm 1, \text{ or } 4D=m^2 \text{ and } U=\pm 2.$$

In fact, the condition $4D = m^2$ cannot arise, since $4ac-4b^2 = m^2$ entails $(2b/m)^2 = -1 \pmod 4$.

The results proved above are summarised in Theorem 7.3:

Theorem 7.3:

Equation (p2) has the solutions $(\pm m, \emptyset)$ irrespective of D.
It has the solutions $(\pm m/2, \pm 1)$ if $D=3m^2/4$,
and the solutions $(\emptyset, \pm 1)$ if $D=m^2$.

The following corollary is obtained by combining Thm.7.3 and Cor.6.1.3:

Cor.7.3.1:

Equation (p1) has the solutions $[\pm 1,\emptyset,\emptyset,\pm 1]$ irrespective of D.
It has the solutions $[+1/2\pm b/m, \pm c/m, \mp a/m, +1/2\mp b/m]$ if $D=3m^2/4$,
and $[-1/2\pm b/m, \pm c/m, \mp a/m, -1/2\mp b/m]$
and the solutions $[\pm b/m, \pm a/m, \mp c/m, \mp b/m]$ if $D=m^2$.

Notes on §7.

Theorems 7.1 and 7.2 are based on [G] Art.'s 171 and 172 respectively. The problem of finding all solutions of $T^2+DU^2 = m^2$ is discussed in Art.179.

Lagarias [L1] has shown that Gauss' algorithm for reduction (as described in the proof of Theorem 7.1) requires $O(M(\log(\|F\|))\log(\|F\|))$ elementary operations and that equivalence of reduced forms (using the criterion of Theorem 7.2) can be decided in $O(M(\log(\|F\|)))$ elementary operations.

All solutions of (p2) are explicitly described in Theorem 7.3, subject only to the determination of m.

§8. The case of positive non-quadratic determinant.

Throughout this section, D will denote a positive non-quadratic integer.

Definition:
A form $F=(a,b,c)$ with determinant D is reduced if
$0 < b < \sqrt{D}$ and $\sqrt{D} - b < |a| < \sqrt{D} + b$. (These two conditions are together equivalent to $|\sqrt{D} - |a|| < b < \sqrt{D}$.)

Lemma 8.1:
Let $F=(a,b,c)$ be a form of determinant D, and suppose that
$$\sqrt{D} - |a| < b < \sqrt{D} \quad \text{and} \quad |c| \geq |a|.$$
Then F is a reduced form.

Proof:
$$\sqrt{D} - a < b < \sqrt{D} \Rightarrow 0 \leq b^2 < \max(D, D+a^2-2|a|\sqrt{D})$$
$$\Rightarrow D + a^2 - b^2 > 0$$
$$\Rightarrow a^2 - ac > 0$$
$$\Rightarrow ac < 0, \text{ since } |c| > |a|.$$
But then $D = b^2 - ac = b^2 + |ac| \geq b^2 + |a|^2$, whence $\sqrt{D} \geq |a|$, and
$$0 < \sqrt{D} - a < b < \sqrt{D}, \text{ and } \sqrt{D} - b < |a| \leq \sqrt{D} < \sqrt{D} + b.$$

Theorem 8.2:
If $F=(a,b,c)$ has determinant D, then F is equivalent to a reduced form.

Proof:
Suppose that F is not itself reduced. Define a sequence of forms
$$F_0 = F, F_1, \ldots, F_t, \ldots$$
where $t \geq 1$ and $F_i = (a_i, b_i, c_i)$, and F_i has matrix M_i, via:
$$M_{i+1} = S_i^T M_i S_i$$
where $S_i = [0,-1,1,k_i]$, and k_i is the unique integer such that
$$\sqrt{D} - |c_i| < -b + k_i c_i < \sqrt{D}.$$
The sequence of positive integers $|a_i|$ defined in this way cannot decrease indefinitely, and there is a first index $r > 0$ such that $|a_{r+1}| \geq |a_r|$. The form F_r will now be shown to be reduced:
A simple computation shows that for $0 \leq i \leq r$
$$a_{i+1} = c_i$$
$$b_{i+1} = -b_i + k_i c_i$$
$$\text{and } c_{i+1} = (b_{i+1}^2 - D)/c_i.$$
The definitions of a_r and b_r ensure that $\sqrt{D} - |a_r| < b_r < \sqrt{D}$, whilst
$$|a_r| \leq |a_{r+1}| = |c_r|.$$
Thus F_r is reduced by Lemma 8.1.

Definition:
The forms $F=(a,b,c)$ and $G=(A,B,C)$ of determinant D are neighbours if
$$A=c \text{ and } B=-b \pmod{c}.$$
G is then "neighbour to F by last part", and F is "neighbour to G by first part". Neighbouring forms are equivalent via a matrix having the form $[0,-1,1,k]$.

Lemma 8.3:
Suppose that $F=(a,b,c)$ is reduced, and has determinant D. Then:
(1) $ac < 0$
(2) $\sqrt{D} - b < |c| < \sqrt{D} + b$, and (c,b,a) is also reduced
(3) if $G=(A,B,C)$, where $A=c$, and $\sqrt{D} - |c| < B = -b + kc < \sqrt{D}$, then G is reduced, and $kc > 0$.
(4) there is exactly one reduced form which is neighbour to F on each side.

Proof:

(1) $D = b^2 - ac$, and $\sqrt{D} > b$, so that $ac < \emptyset$.

(2) $|c| = (D-b^2)/a$, and $\sqrt{D} - b < |a| < \sqrt{D} + b$ ensures that $|c|$ lies in the same interval.

(3) $kc = b+B = (b+\sqrt{D}-|c|) + (B-(\sqrt{D}-|c|)) > \emptyset$, by (2). Thus $kc \geq |c|$, and

$$2B = (\sqrt{D}-b) + (B-(\sqrt{D}-|c|)) + (kc-|c|) > \emptyset,$$

proving that $\emptyset < B < \sqrt{D}$.

Moreover, $\sqrt{D}+B-|c| = (\sqrt{D}-b) + (kc-|c|) > \emptyset$, so that

$$\sqrt{D} - B < |c| = |A| < \sqrt{D} + B.$$

(4) If H is a reduced form which is neighbour to F by last part, then $H = (c,V,W)$, where V and $-b$ are congruent modulo c. But V is uniquely determined, since $\sqrt{D} - V < |c|$, and $V < \sqrt{D}$, and accordingly H is the form G, as defined in (3).

If $H = (r,s,t)$ is neighbour to F by first part, then (t,s,r) is neighbour to (c,b,a) by last part, and is reduced, by (2). But then H itself is also reduced, by (2).

Theorem 8.4:

The set of reduced forms of determinant D is finite, and is a disjoint union of classes ('periods'), each of which consists of a sequence of forms

$$F_0, F_1, \ldots, F_t = F_0,$$

where F_{i+1} is neighbour to F_i by last part for each i.

Moreover, if $M_{i+1} = Q_i^T M_i Q_i$, where M_i is the matrix of $F_i = (a_i,b_i,c_i)$ and Q is the matrix $[\emptyset,-1,1,k_i]$, then in the each of the sequences:

$$a_0, a_1, \ldots, a_t = a_0$$
$$c_0, c_1, \ldots, c_t = c_0$$
$$k_0, k_1, \ldots, k_t = k_0$$

the terms are non-zero integers which alternate in sign. In particular, every period has even length.

Proof:

There are only finitely many integer solutions of $b^2 - ac = D$ such that $ac < \emptyset$, and thus only finitely many reduced forms, by Lemma 8.3 (1).

The reduced forms are therefore distributed into periods as defined, in view of Lemma 8.3 (4).

Lemma 8.3 (1) shows that $a_i c_i < \emptyset$ for each i, so that the sequences of a's and c's consist of non-zero integers alternating in sign. Since $k_i c_i < \emptyset$ for each i by Lemma 8.3 (4), the sequence of k's has a similar property.

The next results will be used to prove that the periods of reduced forms are also the equivalence classes of reduced forms. This is technically the hardest part of Gauss' exposition, and an alternative approach, based upon results of [B], is presented.

Following [B], P (respectively P_1) is used to denote the class of 2x2 integer matrices with non-negative entries and positive determinant (respectively determinant 1). Familiarity with the notation and results of [B] §6 is assumed. The basic result required is the following theorem:

Theorem A:
Let $Z=[p,q,r,s]$ be an integer matrix such that $qr>0$, and $Z[x]=x$ has an irrational root. The set of matrices X in P which commutes with Z is the set of powers of a matrix W, where $w=cap^{-1}(W)$ is the period of the (purely periodic) u-l encoding of the (unique) positive root of $Z[x]=x$.

Proof:
$XZ = ZX \iff X(Z+cI) = (K+cI)X$ for some integer c
$\iff X(Z+cI)^{-1} = (Z+cI)^{-1}X$ for some integer c.
Moreover, $Z[x]=x$ iff $(Z+cI)[x]=x$ iff $(Z+cI)^{-1}[x]=x$.
Since $qr>0$, there is an integer c such that $Z+cI$ or $(Z+cI)^{-1}$ is in P. The result then follows from [B] Theorem 6.4.

Notation:
In the sequel, S will denote the matrix $[0,-1,1,0]$ in P_1.

The relevance of Theorem A is established by the following lemma:

Lemma 8.5:
Let M and X be integer matrices. Then $(-S)X^{-1}S = X^T$, and the following are equivalent:
(A) $X^TMX = M$
(B) $X^{-1}SMX = SM$
and (C) $Y^{-1}MSY = MS$, where $Y = (-S)XS$.

Proof:
Direct calculation shows that $(-S)X^{-1}S = X^T$.
The equivalence of (A),(B) and (C) is an easy consequence of this identity, and the fact that $S^{-1} = -S$.

If $F=(a,b,c)$ is a form of determinant D, and M is the matrix of F, then
$MS = [-b,-c,a,b]$ and $SM = [b,-a,c,-b]$
both satisfy the premises required for the application of Theorem A. Moreover, if v and w are the periodic parts of the encodings of the positive roots of the equations $MS[x]=x$ and $SM[x]=x$ respectively, then $v=rev(int(w))$. (By Lemma 8.5, the matrix X commutes with SM iff X^T commutes with MS. Thus cap(v) is the transpose of cap(w), and $v=rev(int(w))$ by [B] Lemma 5.2 (ii).) These observations have an important role in the following theorem:

Theorem 8.6:

Let F = (a,b,c) be a reduced form of determinant D and matrix M. Let

$$F_0 = F, F_1, \ldots, F_{2r-1} = F_0$$

be the period of F, and define $Q_i = [\emptyset,-1,1,k_i]$ as in Thm.8.4.

If $Q = Q_0 Q_1 \ldots Q_{2r-2} Q_{2r-1}$, then

either

a > ∅, the matrix $\pm Q$ is in P_1, and the u-l encoding of the positive root of SM[x]=x (resp. MS[x]=x) has period

$$q = cap^{-1}(\pm Q) = l^{-k_0}u^{k_1}l^{-k_2}u^{k_3} \ldots l^{-k_{2r-2}}u^{k_{2r-1}} \quad (\text{resp. } rev(int(q)))$$

or

a < ∅, the matrix $\pm Q^{-1}$ is in P_1, and the u-l encoding of the positive root of SM[x]=x (resp. MS[x]=x) has period

$$\bar{q} = cap^{-1}(\pm Q^{-1}) = u^{-k_{2r-1}}l^{k_{2r-2}}u^{-k_{2r-3}}l^{k_{2r-4}} \ldots u^{-k_1}l^{k_0} \quad (\text{resp. } rev(int(q))).$$

Proof:

It is easy to verify that if k is an integer, then

$$L^k S = SU^{-k} = [\emptyset,-1,1,-k] \quad \ldots\ldots\ldots \quad (*)$$

Suppose that a > ∅; then c < ∅, and $k_0 < \emptyset$ by Lemma 8.3. For $1 \leq t \leq r$, let Q(t) denote $Q_0 Q_1 \ldots Q_{2t-2} Q_{2t-1}$. By (*):

$$\begin{aligned} Q(t) &= \prod_{0 \leq i \leq 2r-1} [\emptyset,-1,1,k_i] \\ &= \prod_{0 \leq i \leq r-1} (L^{-k_{2i}}S.L^{-k_{2i+1}}S) \\ &= \prod_{0 \leq i \leq r-1} (-L^{-k_{2i}}(-S).L^{-k_{2i+1}}S) \\ &= (-1)^r \prod_{0 \leq i \leq r-1} L^{-k_{2i}}U^{k_{2i+1}}. \end{aligned}$$

Thus $\pm Q = \pm Q(r)$ is in P_1, and Lemma 8.5 shows that

$$cap^{-1}(\pm Q) = l^{-k_0}u^{k_1}l^{-k_2}u^{k_3} \ldots l^{-k_{2r-2}}u^{k_{2r-1}}$$

is the periodic part of the u-l encoding of the positive root of SM[x]=x.

($cap^{-1}(\pm Q)$ is a concatenation of periods since $Q^T M Q = M$ entails $Q^{-1} SMQ = SM$. It is the minimal period since the periodic part is necessarily of the form Q(t) with $1 \leq t \leq r$, and $(Q(t))^T M Q(t) = M$ only if t=r.)

Suppose that a < ∅; then c > ∅, and $k_0 > \emptyset$ by Lemma 8.3. Then:

$$\begin{aligned} (-S)QS &= (-S) \prod_{0 \leq i \leq 2r-1} [\emptyset,-1,1,k_i].S \\ &= (-S) \prod_{0 \leq i \leq r-1} (SU^{k_{2i}}.SU^{k_{2i+1}}).S \\ &= \prod_{0 \leq i \leq r-1} (-U^{k_{2i}}S.U^{k_{2i+1}}(-S)) \\ &= (-1)^r \prod_{0 \leq i \leq r-1} U^{k_{2i}}L^{-k_{2i+1}}. \end{aligned}$$

Thus, if Y = (-S)QS, then $\pm Y = \pm(Q^{-1})^T$ is in P_1, and moreover

$$v = u^{k_0}l^{-k_1}u^{k_2}l^{-k_3} \ldots u^{k_{2r-2}}l^{-k_{2r-1}}$$

is the periodic part of the u-l encoding of the positive root of MS[x]=x, as before.

Cor. 8.6.1:
Let F=(a,b,c) be a reduced form of determinant D and matrix M, and let w be the period of the u-l encoding of the positive root of SM[x]=x. The form F is uniquely determined by m = gcd(a,2b,c) and w.

Proof:
SM[x]=x iff $ax^2+2bx+c = \emptyset$. The positive root of $ax^2+2bx+c = \emptyset$ is determined by w, in view of the uniqueness of u-l encodings of irrationals. The triple of coefficients (a,b,c) is determined up to an integer multiple by the positive root, and is thus determined up to sign if m is known. But by Theorem 8.6, the sign of a is determined by the first symbol of w.

Notation:
The finite u-l sequence w associated with the form F as in Cor.8.6.1 will be denoted by w(F).

Two finite u-l sequences v and w are said to be "cyclically related" if one is a cyclic permutation of the other.

The next corollary is an easy consequence of Cor.8.6.1 and Thm.8.6.

Cor. 8.6.2:
The forms F = (a,b,c) and G = (A,B,C) are in the same period if and only if $aA > \emptyset$ and w(F) and w(G) are cyclically related, or $aA < \emptyset$ and w(F) and rev(int(w(G)) are cyclically related.

Lemma 8.7:
Let F=(a,b,c) and G=(A,B,C) be forms with matrices M and N respectively, and suppose that Q = [x,y,z,t] is an integer matrix of determinant 1 such that xt is non-zero and $Q^TMQ = N$. Then xt and aA have the same sign.

Proof:
Equating the diagonal entries of Q^TM and NQ^{-1} shows that

$$(B + b)y = Cx - ct \quad \text{and} \quad (B + b)z = At - ax.$$

Multiplying these relations together then proves that

$$(B + b)^2 yz - (AC + ac)xt = - aCx^2 - Act^2 \quad \ldots\ldots\ldots \quad (+).$$

Since xt-yz=1, where xt is non-zero, and AC and ac are negative, the left-hand side of (+) has the same sign as xt. Since AC and ac are negative, aC and Ac necessarily have the same sign. But this is consistent with (+) only if aC and xt have opposite signs, that is, only if aA and xt have the same sign.

Theorem 8.8:
The periods are the equivalence classes of reduced forms of determinant D.

Proof:
Let F=(a,b,c) and G=(A,B,C) be reduced forms of determinant D which are equivalent. It will suffice to show that F and G lie in the same period.

Let M and N be the matrices of the forms F and G respectively, and let Q = [x,y,z,t] be an integer matrix of determinant 1 such that $Q^TMQ = N$.

If x or t = Ø, then F and G are neighbouring forms, and lie in the same period by Lemma 8.3. The case x = t = Ø cannot arise, as F and G cannot both be reduced if Q = ±S.

Suppose then that xt is non-zero. If xt > Ø, then Q or Q^{-1} is in P (modulo sign, which may be arbitrarily chosen). Since Q^{-1} defines an equivalence between G and F, the case "Q is in P " can be considered without loss of generality.

Let v = w(F) and w = w(G), and let <v> and <w> be the encodings of the irrationals f and g respectively. Since $Q^TMQ = N$ implies $Q^{-1}SMQ = SN$, and g is the positive root of SN[x] = x, it follows that Q[g] is the positive root of SM[y] = y; that is Q[g] = f. Equating the u-l encodings of Q[g] and f, it follows that <v> = q.<w>, where $q = cap^{-1}(Q)$, whence v and w are necessarily cyclically related. Thus F and G are in the same period, by Lemma 8.7 and Cor.8.6.2.

Now suppose that xt < Ø. Then either -SQ or QS (modulo sign, as before) is in P_1. If -SQ is in P_1, then

$$Q^TMQ = N \Rightarrow Q^{-1}SMQ = SN \Rightarrow (-SQ)^{-1}MS(-SQ) = SN,$$

whence the positive root of MS[x] = x is (-SQ)[g]. Equating u-l encodings then proves that w and rev(int(v)) are cyclically related, so that F and G are in the same period by Lemma 8.7, and Cor.8.6.2.

If QS is in P_1, then the equation $(QS)^{-1}SM(QS) = NS$ may be used in a similar fashion to prove that v and rev(int(w)) are cyclically related, and F and G are in the same period.

Remark:
The proof of Thm.8.8. can be abbreviated; it suffices to prove that F and G lie in the same period, so that (by replacing F and/or G by a neighbouring reduced form if necessary) both a and A can be supposed positive. Lemma 8.7 then shows that only the case xt > Ø can arise.

The proof as given preserves the symmetry between the cases aA < Ø and aA > Ø.

Finding all solutions of $T^2 - DU^2 = m^2$.

Let $F = (a,b,c)$ be a form of determinant D, and let $m = \gcd(a,2b,c)$. By Cor.6.1.3, solutions of $T^2 - DU^2 = m^2$ are in 1-1 correspondence with proper equivalences between the form $F = (a,b,c)$ and itself. Explicitly, if (T,U) is a solution of $T^2 - DU^2 = m^2$, and

$$x = (T+bU)/m,\quad y = cU/m,\quad z = (-aU)/m \quad \text{and} \quad t = (T-bU)/m$$

then $X = [x,y,z,t]$ defines a proper equivalence between F and itself.

All solutions of $T^2 - DU^2 = m^2$ are readily derived from the set of non-negative solutions. If T and U are non-negative integers, then the coefficients y and z of the associated matrix X above have the same sign as c (or equivalently -a), whilst

$$xt = (T^2 - b^2U^2)/m^2 = (T^2 - DU^2 - acU^2)/m^2 = (m^2 - acU^2)/m^2,$$

so that x and t are both positive. Hence $X = [x,y,z,t]$ is in P_1 if a is positive, and X^{-1} is in P_1 if a is negative. In either event, if Y denotes whichever of X and X^{-1} is in P_1, then $Y^T SMY = SM$ by Lemma 8.5. By [B], Theorem 6.4, it then follows that Y is of the form W^K, where $\mathrm{cap}^{-1}(W)$ is the period of the u-l encoding of the positive root of $SM[x]=x$. This proves:

Theorem 8.9:

Let F and Q be as in Theorem 8.6.

The matrix X defines a proper equivalence between F and itself if and only if X is of the form Q^k, where k is an arbitrary integer.

Cor. 8.9.1:

Let Q and F be as in Theorem 8.6, and let $m = \gcd(a,2b,c)$. Let Y denote whichever of Q and Q^{-1} is in P_1.

If Y^K is the matrix $[x_k, y_k, z_k, t_k]$, then

$$(T_k, U_k) = \left(m(x_k+t_k)/2 \ , \ m(x_k-t_k)/2b \right)$$

is the general solution of $T^2 - DU^2 = m^2$ in non-negative integers, and the sequences T_k and U_k are defined by a common linear recurrence

$$Z_{k+1} = (2T_1/m).Z_k - Z_{k-1},$$

from $T_0 = m$, $U_0 = 0$, and the least positive solution (T_1, U_1).

In particular, if $v_+ = (T_1 + U_1\sqrt{D})/m$ and $v_- = (T_1 - U_1\sqrt{D})/m$, then

$$T_k = (m/2).(v_+^K + v_-^K), \text{ and } U_k = (m/(2\sqrt{D})).(v_+^K - v_-^K).$$

Proof:

The form of the general solution of $T^2 - DU^2 = m^2$ can be simply derived from Cor.'s 6.1.1 and 6.1.3.

Let T,U,x,y,z and t respectively denote T_1, U_1, x_1, y_1, z_1 and t_1. Then, using the relations between T,U,x,y,z and t given by Cor.6.1.1:

$$\begin{aligned} T_{k+1} &= m(x_{k+1}+t_{k+1})/2 \\ &= m(xx_k+yz_k+zy_k+tt_k)/2 \quad \text{(from the relation } Y^{k+1} = Y.Y^k\text{)} \\ &= ((T+bU)x_k+cUz_k-aUy_k+(T-bU)t_k)/2 \\ &= (2T/m)(m(x_k+t_k)/2) - ((T-bU)x_k-cUz_k+aUy_k+(T+bU)t_k)/2 \\ &= (2T/m).T_k - m(tx_k-yz_k-zy_k+xt_k)/2 \\ &= (2T/m).T_k - T_{k-1} \quad \text{(from the relation } Y^{k-1} = Y^{-1}.Y^k\text{)}. \end{aligned}$$

The same linear recurrence relation can be derived for the U's in a similar fashion.

The explicit solution of the linear recurrences then yields the 'closed form' for T_k and U_k as described.

Notes on §8.

Theorem 8.2, Lemma 8.3 and Theorem 8.4 are based on [G] Art.'s 183-7. Theorems 8.6 and 8.8 are respectively associated with Art.191 and 193, but Gauss avoids explicit use of continued fractions. Apart from the unusual notation for continued fractions, the latter part of this section is similar to Dirichlet's approach in [D] (see Smith [Sm] Art.93).

The solution of $T^2 - DU^2 = m^2$ is discussed in Art.198.

Gauss' reduction procedure (as described in the proof of Theorem 8.2) may require $O(\|F\|^{1/4})$ elementary operations (see [L1]). An alternative reduction procedure, which terminates in $O(M(\log(\|F\|))\log(\|F\|))$ elementary operations, is described by Lagarias [L1].

Theorem 8.8 is the basis for a naive algorithm for deciding equivalence of reduced forms of determinant D. The number of reduced forms in a period is the number of terms in the period of the classical continued fraction expansion of an appropriate quadratic irrational (see Theorem 8.6). For instance, the period of a suitable reduced form may have the same length as the period of the continued fraction expansion of $\sqrt{D}$, which (in certain cases) exceeds $\sqrt{D}(\log(D))^{-1}/3$ (see [L1] and [L2]). There is no known procedure for deciding equivalence of reduced forms of determinant D which terminates in time bounded by a polynomial in log(D).

Similarly, the computation of the minimal solution of $T^2-DU^2 = m^2$ (determining Q as defined in Theorem 8.6, and applying Theorem 8.9) may require O(D) elementary operations, and, in general, the least non-trivial solution (T,U) cannot even be written down in time bounded by a polynomial in log(D) (see Shanks [Sh] and Lagarias [L1],[L2]).

§9. The case of quadratic determinant.

Throughout this section, D will denote a quadratic integer h^2.

Definition:
A form F=(a,b,c) with determinant $D = h^2$ is reduced if
(1) $\emptyset \leqslant a \leqslant 2h-1$
(2) b=h
and (3) $c=\emptyset$.

Theorem 9.1:
Let F be a form of determinant D. There is a reduced form $(a,h,\emptyset)$ equivalent to F.

Proof:
Since $h^2 - b^2 = ac$, there are co-prime integers p and q such that
$$(h-b)/a = -c/(h+b) = q/p.$$
Suppose that P and Q are chosen so that $Pp + Qq = 1$, and let X be the matrix $[P,q,Q,p]$. Let M be the matrix of F. Direct calculation shows that $N=X^TMX$ is the matrix of the form G, where G = (A,B,C), and
$$B = Pqa + (Pp+Qq)b + Qpc = Pp(h-b) + b + Qq(h+b) = h,$$
$$C = q^2a + 2pqb + p^2c = pq(h-b) + 2pqb - pq(h+b) = \emptyset.$$
Now if $Z = [1,\emptyset,k,1]$, then Z^TNZ is the matrix of the form $(A+2Bk+Ck^2, B+Ck, C) = (A+2hk, h, \emptyset)$, and an appropriate choice of k will ensure that $\emptyset \leqslant A+2hk \leqslant 2h-1$.

Theorem 9.2:
Two reduced forms of determinant D are equivalent if and only if they are equal.

Proof:
Suppose that $F=(r,h,\emptyset)$ and $G=(s,h,\emptyset)$ have matrices M and N respectively, and let $X=[x,y,z,t]$ be an integer matrix of determinant 1 such that $X^TMX = N$. Then $X^TM = NX^{-1}$, and by equating coefficients:
$$rx+hz = st-hz \text{ and } -sy = ry = 2hy = \emptyset.$$
Thus $y=\emptyset$, and $st-rx = 2hz$. But then $xt = 1$, so that $x = t = \pm 1$, and
$$|2hz| = |rx-st| = |r-s| \leqslant 2h-1.$$
Thus $z=\emptyset$, and r=s.

Theorems 9.1 and 9.2 solve the problem of deciding equivalence of forms of determinant D, and of determining an explicit equivalence when appropriate. It remains to consider the problem of determining all equivalences between two forms of determinant D. For this purpose, the result of Cor.6.1.3 is used in conjunction with the following lemma:

Lemma 9.3:
Let F=(a,b,c) be a form of determinant D, and let m = gcd(a,2b,c). If t and u are integers such that $t^2 - Du^2 = m^2$, then $t = \pm m$ and $u = \emptyset$.

Proof:
Suppose that t and u satisfy $t^2 - Du^2 = m^2$. By Cor.6.1.1, m divides 2t, and the relation
$$(2t/m)^2 - (2h/m)^2 u^2 = (2t/m-2hu/m)(2t/m+2hu/m) = 4$$
entails $2t/m = \pm 2$, and $u = \emptyset$.

As remarked in §5, the above methods are sufficient to solve Problem 1 under the hypothesis that K is a non-zero integer. (Thm.5.2 requires K to be non-zero.) An alternative direct method for solving Problem 1 for general K will now be described.

As in the proof of Thm.9.1, there are co-prime integers p and q with

$$(h-b)/a = -c/(h+b) = q/p.$$

On defining $f = (h-b)/q$ and $g = -(h+b)/p$ (which are integers), direct calculation verifies that $ax^2 + 2bxy + cy^2$ factorises as $(px-qy)(fx-gy)$. There are then two cases to consider:

Case 1: $K = 0$

Then $ax^2 + 2bxy + cy^2 = K = 0$ if and only if either $px=qy$ or $fx=gy$. If $d=\gcd(u,v)$, then the general solution of $ux=vy$ is

$$x(k) = (v/d)k, \quad y(k) = (u/d)k, \text{ where } k \text{ is an integer.}$$

Case 2: $K \neq 0$

Then $ax^2 + 2bxy + cy^2 = (px-qy)(fx-gy) = K$ if and only if $K = rs$ and x and y are integers which simultaneously satisfy $px-qy = r$ and $fx-gy = s$.

Since $pg-qf = -2h$, this pair of linear equations is non-singular and has a unique rational solution (x,y) for each factorisation of K as rs. Thus, to obtain all solutions of Problem 1 in this case, it suffices to eliminate non-integral pairs from the finite set of rational pairs (x,y) derived from all possible factorisations of K in this way.

Notes on §9.

Theorem 9.1, Theorem 9.2 and Lemma 9.3 are based on [G] Art.'s 206, 207 and 209 respectively. The alternative method of solution is described in Art.212.

Gauss' reduction procedure (as described in the proof of Theorem 9.1) requires at most $O(M(\log(\|F\|))\log(\|F\|))$ elementary operations. Equivalence of reduced forms can be decided in $O(M(\log(D)))$ elementary operations (see Lagarias [L1]).

When K is non-zero, the alternative direct method of solution of Problem 1 has the disadvantage that factorisation of K is required.

§10. Solving a general quadratic equation in integers.

Problem 2 is concerned with solving a general quadratic equation in integers; that is, finding all integer pairs (x,y) such that

$$ax^2 + 2bxy + cy^2 + 2dx + 2ey + f = 0 \quad \ldots\ldots\ldots \text{(E1)},$$

where a,b,c,d,e and f are given integers. Provided that $D = b^2 - ac = 0$, the substitutions $p = Dx + be - cd$ and $q = Dy + bd - ae$ transform the above problem to finding all solutions of

$$ap^2 + 2bpq + cq^2 = K \quad \ldots\ldots\ldots \text{(E2)}$$

where $-K$ is the integer $f(b^2 - ac) + (b^2 - ac)(ae^2 - 2bde + cd^2)$, and the associated values of x and y, viz:

$$x = (p + cd - be)/D \quad \text{and} \quad y = (q + ae - bd)/D \quad \ldots\ldots \text{(\#)},$$

are integral. When the set of possible solutions of (E2) is known to be finite (as for example, when $D<0$), all integer solutions of (E1) can be effectively identified by enumerating the pairs (x,y) defined by (#). If, on the other hand, (E2) has infinitely many solutions, special techniques are required to identify those values of p and q - if any - which give rise to integral values of x and y in (#). In particular, it is possible for (E1) to have no integer solution even though (E2) admits an infinity of integer solutions.

From previous results, equation (E2) admits an infinity of solutions in just two cases: the case of positive non-quadratic determinant D, and the case when D is positive quadratic, and K is 0.

Case 1: D is positive quadratic, and K is 0.

In this case, the general solution of (E2) has the form

$$p = Az, \quad q = Bz,$$

where A and B are co-prime and z is an arbitrary integer. If D=1, all the solutions of (#) derived from solutions (p,q) of (E2) will be integral, but there is otherwise a possibility of rational solutions.

Let p=Az, q=Bz, and let P and Q be integers such that $PA + QB = 1$. If x and y are defined as in (#), then

$$(Px + Qy).D = z + P(cd-be) + Q(ae-bd),$$

so that a necessary condition for x and y to be integral is that

$$z = P(cd-be) + Q(ae-bd) \pmod{D}.$$

Substituting $z = k.D + P(cd-be) + Q(ae-bd)$ and p=Az in (#) then shows that all integral values of x and y satisfying (E1) are of the form:

$$x = Ak + Qw/D, \quad y = Bk - Pw/D,$$

where $w = A(bd-ae)-B(be-cd)$. A necessary and sufficient condition for one pair (x,y) of this form to be integral is that D should divide Pw and Qw. Since P and Q are co-prime, this is the case if and only if D divides w, and (#) yields an integral solution of (E1) if and only if all solutions given by (#) are integral. To summarise:

The decision procedure.

Suppose that $(p,q) = (Az,Bz)$ is the general solution of (E2). Determine $w = A(bd-ae)-B(be-cd)$. If D divides w, then all pairs (x,y) derived from (#) will be integral solutions of (E1), but otherwise (1) has no integral solution.

Case 2: D is positive non-quadratic.

In order to devise a decision procedure in this case, some additional results are required.

Lemma 10.1:
Let $D=n^2D'$, where D' is square-free, and let m be a positive integer.
There is a form F=(a,b,c) of determinant D such that gcd(a,2b,c) = m
if and only if either $D' = 1 \pmod 4$ and m divides 2n
or $D' = 2,3 \pmod 4$ and m divides n.

Proof:
Sufficiency: If $D'=1 \pmod 4$ and m divides 2n, then the form
$$F = (m,n,n^2(1-D')/m)$$
has determinant D, and m divides both 2n and $n^2(1-D')/m$.
If $D'=2,3 \pmod 4$ and m divides n, then the form
$F = (m,0,-n^2D'/m)$ has determinant D, and m divides $-n^2D'/m$.
Necessity: Let F = (a,b,c) be a form of determinant D such that gcd(a,2b,c) = m.
Irrespective of the residue of D', the integer $4n^2D' = 4D = 4(b^2-ac)$ is divisible by m^2. Let g = gcd(m,2n), and suppose that m=Mg and 2n=Ng. Now M and N are co-prime, and M^2 divides N^2D'. Thus M^2 divides D', which is square-free, and M=1. Hence m divides 2n.
Now suppose that $D' = 2,3 \pmod 4$. In view of the relation
$$(2n/m)^2 D' = (2b/m)^2 - 4ac/m^2 = (2b/m)^2 \pmod 4,$$
and the fact that D' is not a quadratic residue modulo 4, the quotient 2n/m is even, and m divides n in this case.

Lemma 10.2:
Let F, m and the sequences of integers T_k and U_k be as in Cor.8.9.1.
For each integer N > 1, there is an integer R such that
$$T_{k+R} = T_k \pmod N \text{ and } U_{k+R} = U_k \pmod N$$
for all integers $k \geq 0$, and R can be calculated without explicitly determining T_k and U_k for $0 \leq k \leq R$.
In particular, the sequences of residues $T_k \pmod N$ and $U_k \pmod N$ are purely periodic.

Proof:
Let (T,U) denote (T_1,U_1).
In view of the recurrence relations which define the sequences T_k and U_k, (see Cor.8.9.1), it will suffice to prove that for some index R:
$$T_R = T_0 = m \pmod N,\ U_R = U_0 = 0 \pmod N,$$
$$T_{R+1} = T \pmod N \text{ and } U_{R+1} = U \pmod N.$$
By Lemma 10.1, the existence of the form F=(a,b,c) of determinant D such that gcd(a,2b,c) = m is sufficient to ensure the existence of a form G=(A,B,C) of determinant N^2D such that gcd(A,2B,C) = m. This ensures that the equation $T^2 - N^2DU^2 = m^2$ has a solution (T*,U*) in positive integers, in view of Cor.8.9.1. Moreover, since all such solutions of $T^2 - DU^2 = m^2$ are of the form (T_k,U_k), there is an index r such that $(T_r,U_r) = (T^*,NU^*)$. (Note that r can be in principle be computed from the explicit form for T_k given in Cor.8.9.1.) For this r, the required conditions are satisfied in part, since $U_r = 0 \pmod N$. It may be that T_r, T_{r+1} and U_{r+1} also satisfy the appropriate congruences modulo N, but in any event it can be shown that both sets of conditions are satisfied for the index R=2r. Explicitly
$$T_k = (m/2).(v_+^k + v_-^k) \text{ and } U_k = (m/(2\sqrt{D})).(v_+^k - v_-^k),$$
where $v_+ = (T + U\sqrt{D})/m$ and $v_- = (T - U\sqrt{D})/m$. Thus, noting that m divides $2T_k$ by Cor.6.1.1, that m divides 2D by Lemma 10.1, and that $v_+v_- = 1$, and using the relations above:
$$T_R = (1/m).(T_r^2+DU_r^2) = (1/m).(m^2+2DU_r^2) = m+(2D/m)U_r^2 = m \pmod N,$$
$$U_R = (2T_r/m).U_r = 0 \pmod N,$$
$$T_{R+1} = (2D/m)U_rU_{r+1} + T = T \pmod N$$
$$U_{R+1} = (2T_{r+1}/m)U_r + U = U \pmod N.$$

The decision procedure.

By Cor.6.1.5 the general solution of equation (E2) has the form

$$p = (1/m).(A(\pm T_k) + B(\pm U_k)),\ q = (1/m).(C(\pm T_k) + D(\pm U_k)),$$

which can be conceived as four families of solutions of the form

$$p = (1/m).(AT_k + BU_k),\ q = (1/m).(CT_k + DU_k)$$

by appropriately choosing the signs of the integers A,B,C and D. For each such family, the problem of deciding whether the associated solution (x,y) of (E1), as defined by (#), is integral, depends only upon the residues of T_k and U_k modulo mD. Taking N=mD in Lemma 10.2, and computing the index R as prescribed in the proof of Lemma 10.2 then ensures that all possible residues of T_k and U_k modulo N are represented amongst the pairs (T_k,U_k), where $0 \leqslant k < R$. Moreover, (T_i,U_i) yields an integral solution of (E1) if and only if (T_j,U_j) yields an integral solution, where $j = i \pmod R$.

Notes for §10.

The general principles used in reducing Problem 2 to Problem 1 are based on [G] Art.216. The decision procedures discussed under Cases 1 and 2 are given in Art.'s 218 and 217 respectively. Lemmas 10.1 and 10.2 are based on Art.201; the confusion of suffices and exponents in this article (as it appears in Clarke's English translation) presented special problems here!

References.

[B] On Raney's binary encoding for continued fractions, etc.
W.M.Beynon Theory of Computation Report #34, Univ. of Warwick (1981).

[D] Vereinfachung der Theorie der binaren quadratische Formen.
L.Dirichlet Memoirs of the Academy of Berlin, 1854.

[G] Disquisitiones Arithmeticae, 1801; English translation.
C.F.Gauss (trans. A.A.Clarke) Yale U.P., 1966.

[L1] Worst-case complexity bounds for algorithms in the theory of integral quadratic forms.
J.C.Lagarias Journal of Algorithms 1, 142-207 (1980).

[L2] On the computational complexity of determining the solvability or unsolvability of the equation $X^2 - DY^2 = -1$.
J.C.Lagarias Trans.Amer.Math.Soc. 260(2), 485-507, 1980.

[Sh] Five number-theoretic algorithms.
D.Shanks Proc. 2nd Manitoba Conf. on Numerical Math. (1972). 51-70.

[Sm] Collected mathematical papers, Vol.1.
H.J.S.Smith Chelsea, New York (1965).